MycoKeys 13: 1–20 (2016)
doi: 10.3897/mycokeys.13.8140
http://mycokeys.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
MycoKeys
Launched to accelerate biodiversity research

Genomics and metagenomics technologies to recover ribosomal DNA and single-copy genes from old fruit-body and ectomycorrhiza specimens

Leho Tedersoo[1], Ingrid Liiv[2], Paula Ann Kivistik[3], Sten Anslan[2], Urmas Kõljalg[2], Mohammad Bahram[2,4]

1 *Natural History Museum, University of Tartu, 14A Ravila, 51005 Tartu, Estonia* **2** *Institute of Ecology and Earth Sciences, University of Tartu, 14A Ravila, 51005 Tartu, Estonia* **3** *Estonian Genome Center, University of Tartu, Riia 23b, 51010 Tartu, Estonia* **4** *Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden*

Corresponding author: *Leho Tedersoo* (leho.tedersoo@ut.ee)

Academic editor: *T. Lumbsch* | Received 15 February 2016 | Accepted 3 April 2016 | Published 13 May 2016

Citation: Tedersoo L, Liiv I, Kivistik PA, Anslan S, Kõljalg U, Bahram M (2016) Genomics and metagenomics technologies to recover ribosomal DNA and single-copy genes from old fruit-body and ectomycorrhiza specimens. MycoKeys 13: 1–20. doi: 10.3897/mycokeys.13.8140

Abstract

High-throughput sequencing (HTS) has become a standard technique for genomics, metagenomics and taxonomy, but these analyses typically require large amounts of high-quality DNA that is difficult to obtain from uncultivable organisms including fungi with no living culture or fruit-body representatives. By using 1 ng DNA and low coverage Illumina HiSeq HTS, we evaluated the usefulness of genomics and metagenomics tools to recover fungal barcoding genes from old and problematic specimens of fruit-bodies and ectomycorrhizal (EcM) root tips. Ribosomal DNA and single-copy genes were successfully recovered from both fruit-body and EcM specimens typically <10 years old (maximum, 17 years). Samples with maximum obtained DNA concentration <0.2 ng µl-1 were sequenced poorly. Fungal rDNA molecules assembled from complex mock community and soil revealed a large proportion of chimeras and artefactual consensus sequences of closely related taxa. Genomics and metagenomics tools enable recovery of fungal genomes from very low initial amounts of DNA from fruit-bodies and ectomycorrhizas, but these genomes include a large proportion of prokaryote and other eukaryote DNA. Nonetheless, the recovered scaffolds provide an important source for phylogenetic and phylogenomic analyses and mining of functional genes.

Key words

Fungal fruit-bodies, low-coverage genome reconstruction, metagenome analysis, functional gene mining, Illumina HiSeq

Copyright *Leho Tedersoo et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Introduction

DNA sequences of high quality are essential for precise molecular identification of organisms and construction of phylogenies. For these purposes, inclusion of type specimens of the species is of utmost importance, because they carry taxonomic information and anchor the target species amongst potentially multiple cryptic taxa (Federhen 2014). However, type specimens of most taxa are decades or even centuries old and their DNA is often poorly preserved due to unsuitable storage conditions such as high humidity and temperature, insufficient care, etc. Therefore, extraction of high-quality DNA as well as amplification and sequencing from old material is painstaking and often virtually impossible (Pääbo et al. 2004).

Fungi represent one of the most diverse groups of eukaryotes with potentially millions of species and a high incidence of sympatric and allopatric cryptic species (Blackwell 2011). Both fruit-bodies and living cultures may serve as type specimens and form a basis for morphological, biochemical and molecular species recognition. However, the vast majority of fungi form no fruit-bodies and cannot be cultured with available techniques. Molecular identification methods have shed light into the high and undescribed fungal diversity in complex substrates such as roots, soil, sediments, water and foliage that are not represented by sequenced material from specimens (O'Brien et al. 2005; Jones et al. 2011; Tedersoo et al. 2014).

Other co-occurring organisms in voucher specimens may hamper molecular identification and genomic analyses of the target specimen. In living cultures, only endohyphal bacteria are common, but fruit-bodies are often infested with prokaryotes, protists, other fungi and meiofauna (nematodes, collembolans, Diptera larvae, etc.). Ectomycorrhizal (EcM) root tips and lesions on plant leaves are usually dominated by a single causal biotroph, although a vast diversity of microscopic organisms co-occurs (Tedersoo et al. 2009; Yoshida et al. 2013). Because of senescence and other resident taxa, certain fungal species and a substantial fraction (up to 5%) of distinct EcM morphotypes consistently remain unsequenced using the combination of fungal and universal primers and Sanger sequencing (Tedersoo et al. 2008; Nguyen et al. 2013).

DNA sequences from the nuclear ribosomal RNA cistron have been widely used, both for identification and phylogenetics of fungi due to a large number of copies and the level of conservation sufficient for discriminating between individuals (the intergenic spacer; IGS – Guidot et al. 1999), species (the internal transcribed spacer; ITS – Gardes et al. 1991; Kõljalg et al. 2005; Schoch et al. 2012) and higher taxa (large subunit; LSU and small subunit; SSU – Gueho et al. 1989). While the nuclear rDNA is distributed in tens to a few hundred tandem repeats (Baldrian et al. 2013), mitochondrial DNA is also abundant due to the presence of multiple mitochondria in active cells that render both targets easy to amplify and use for phylogenetics and identification purposes. Certain single-copy genes (SCGs) such as Translation Elongation Factor 1 α (TEF1) and RNA Polymerase II subunits (RPB1, RPB2) frequently serve to improve phylogenetic resolution, although their amplification and sequencing may require extra care (Schoch et al. 2012). The amplified size of these markers typically range from 300 to 1500 bases, although both LSU and phylogenetically informative single-copy genes are much longer.

The rationale for using such medium-size fragments is the ease of amplification and the ability of Sanger sequences to cover 1000 bases with high quality.

The rapid development of high-throughput sequencing (HTS) tools has greatly improved our understanding about the phylogeny, genome structure and functioning of fungi (Martin et al. 2008; Dentinger et al. 2016; Kohler et al. 2015). Although the HTS genomics approach (i.e., genome-wide sequencing of a single target organism) is commonly used on living cultures, it also enables to incorporate molecular data from herbarium collections of infected plant leaves and old specimens with degraded DNA (Staats et al. 2013; Yoshida et al. 2013; Dentinger et al. 2016). Single nucleotide polymorphisms (SNPs) and phylogenetically informative marker genes can be rigorously extracted from these genomes and used for phylogenetic reconstruction at the level of isolates to kingdoms (Liti et al. 2010; Capella-Gutierrez et al. 2012; Dentinger et al. 2016). These genomic studies have targeted >100-fold coverage that enables very high accuracy but restricts analysis to a few specimens in a single HTS run. As opposed to genomics, 'metagenomics' is a term for untargeted genome-wide sequencing of all organisms in a sample. This approach is mostly used to study the gene content of environmental samples, but sequencing at the depth of hundreds of millions of reads allows to separate nearly full genomes of the dominant prokaryote taxa (Wrighton et al. 2012).

Using Illumina HiSeq 2x150 paired-end sequencing technology, we evaluate the usefulness of low-coverage genomics and metagenomics analyses for recovering barcoding and other phylogenetically informative genes from voucher specimens of fruit-bodies and mycorrhizas in 85 samples simultaneously. In particular, we aimed to i) develop a protocol for genomics and metagenomics from minute amounts of material; ii) evaluate the possibility to obtain high-quality rDNA and SCG sequence data from old type specimens and root tips; and iii) explain why fruit-bodies and EcM root tips of certain taxa consistently fail to amplify and sequence. The ultimate purpose of this study is to extend the public record of high-quality DNA sequences from taxonomically valuable fruit-body voucher specimens and EcM fungal lineages.

Methods

Specimens

For genomics analysis, we selected 56 voucher specimens of fruit-bodies collected from all continents within the last 54 years (Table 1). These specimens are deposited in the fungaria of Tartu University (TU) and Estonian University of Life Sciences (TAA), with a few additional specimens representing loans from the Plant Pathology herbarium of New South Wales, Australia (DAR). We paid particular attention to cover i) old specimens including holotypes (category 'old': n=21; median age, 17.5 years; range, 10.2–53.5 years since the analysis in January, 2015), ii) species with minute-sized (apothecial Helotiales, sequestrate Endogonales) or corticioid (Thelephorales, Atheliales) fruit-bodies that are all inherently exposed to external contamination ('regular': n=19; median age, 5.6 years; range, 2.0–8.8 years, and iii) species that have consistently failed

Table 1. Fruit-body specimens used for genomic sequencing analysis.

Herbarium code	Identification, EcM lineage	Category	Collection date	Biosample
DAR69412	*Densospora nuda* (holotype)	Old	1989-08-19	SAMN04578188
DAR69419	*Densospora nanospora* (holotype)	Old	1989-08-31	SAMN04578189
DAR69421	*Densospora solicarpa* (holotype)	Old	1989-08-31	SAMN04578190
DAR69441	*Endogone magnospora* (holotype)	Old	1991-09-25	SAMN04578191
TAAM 042608	*Rutstroemia juglandis* (holotype)	Old	1961-xx-xx	SAMN04578222
TAAM 137803	*Sarconiptera vinacea* (holotype)	Old	2000-xx-xx	SAMN04578223
TAAM 159500	*Pseudotomentella atrofusca*	Old	1996-09-03	SAMN04578235
TAAM 166877	*Tomentella ferruginea*	Old	1997-08-18	SAMN04578245
TAAM 181146	*Bankera violascens*	Old	2001-09-25	SAMN04578233
TAAM 182408	*Larissia pyrola* (holotype)	Old	1980-xx-xx	SAMN04578220
TAAM 190020	*Arctomollisia kolymensis* (holotype)	Old	1975-xx-xx	SAMN04578221
TAAM 194916	*Lasiomollisia phalaridis* (holotype)	Old	2003-xx-xx	SAMN04578224
TU100021	*Pseudotomentella sp. nov.*	Old	2004-11-03	SAMN04578243
TU100364	*Odontia cf. fibrosa*	Regular	2006-08-04	SAMN04578228
TU100621	*Amaurodon mustialaensis*	Regular	2006-09-28	SAMN04578251
TU100663	*Sarcodon squamosus*	Regular	2006-10-06	SAMN04578240
TU105081	Thelephorales, *Fam. nov.*	Regular	2006-03-05	SAMN04578226
TU108047	*Pseudotomentella mucidula*	Regular	2008-08-27	SAMN04578242
TU108089	*Phellodon tomentosus*	Regular	2008-09-10	SAMN04578241
TU108144	*Tomentellopsis echinospora*	Regular	2008-09-27	SAMN04578250
TU108291	*Tomentella sp. nov.*	Regular	2009-05-01	SAMN04578247
TU108357	*Pseudotomentella armata, comb.ined*	Regular	2009-05-08	SAMN04578246
TU108377	*Thelephora terrestris*	Regular	2009-08-26	SAMN04578229
TU108482	Thelephorales, *Fam. nov.*	Regular	2010-03-17	SAMN04578248
TU110716	Ceratobasidiaceae, /ceratobasidium1	Regular	2011-12-06	SAMN04578167
TU110838	Thelephorales, *Fam. nov.*	Regular	2012-09-24	SAMN04578168
TU113361	*Endogone*	Unseq.[1]	2014-09-27	SAMN04578192
TU115221	Thelephorales, *Fam. nov.*	Regular	2009-10-19	SAMN04578249
TU115235	Thelephorales, *Fam. nov.*	Old	1997-06-12	SAMN04578230
TU115270	*Pseudotomentella italica, comb.ined.*	Regular	2008-08-09	SAMN04578244
TU115333	*Boletopsis leucomelaena*	Regular	2011-09-09	SAMN04578187
TU115426	Thelephorales, *Fam. nov.*	Regular	2012-08-28	SAMN04578172
TU116148	Atheliales; /atheliales1	Regular	2013-01-14	SAMN04578173
TU116208	*Cantharellus*	Unseq.	2013-07-15	SAMN04578174
TU116326	*Helvella*	Unseq.	2013-09-19	SAMN04578175
TU116380	*Helvella*	Unseq.	2013-10-13	SAMN04578176
TU116400	*Helvella*	Unseq.	2013-11-16	SAMN04578177
TU116448	Pezizaceae	Unseq.	2014-08-09	SAMN04578178
TU116491	*Helvella*	Unseq.	2014-08-11	SAMN04578169
TU116505	*Hydnum*	Unseq.	2014-08-11	SAMN04578179
TU116506	*Cantharellus*	Unseq.	2014-08-11	SAMN04578180
TU116517	*Helvella*	Unseq.	2014-08-11	SAMN04578181
TU116528	*Clavulina*	Unseq.	2014-08-12	SAMN04578182
TU116531	*Helvella*	Unseq.	2014-08-12	SAMN04578171
TU116607	*Coltricia*	Unseq.	2014-08-12	SAMN04578183
TU116615	*Helvella*	Unseq.	2014-08-12	SAMN04578171
TU116680	*Endogone*	Unseq.	2014-10-20	SAMN04578184
TU116699	*Glomus macrocarpum*	Unseq.	2014-10-21	SAMN04578185

Herbarium code	Identification, EcM lineage	Category	Collection date	Biosample
TU118650	*Hydnellum ferrugineum*	Regular	2012-08-28	SAMN04578186
TU115206	*Pseudotomentella humicola*	Old	1997-xx-xx	SAMN04578231
TU123535	*Lenzitopsis oxycedri*	Old	1991-04-26	SAMN04578232
TU100990	*Tomentella subamyloidea* (isotype)	Old	1999-08-24	SAMN04578234
FP133500	*Pseudotomentella fumosa* (holotype)	Old	1972-11-16	SAMN04578236
FP133849	*Pseudotomentella molybdea* (holotype)	Old	1974-11-06	SAMN04578237
FP134609	*Pseudotomentella kaniksuensis* (holotype)	Old	1981-07-23	SAMN04578238
SSMF695-4961	*Pseudotomentella griseopergamacea* (holotype)	Old	1961-10-21	SAMN04578239

[1]Unseq., unsequenced

to amplify or sequence in spite of using different primers and targeting different rDNA regions ('unsequenced': n=16; median age, 0.4 years; range, 0.2–1.5 years; recent collections were used to rule out potentially confounding storage effects). Notably, the 'old' specimens were comprised mainly of Thelephorales, Helotiales and Endogonales, whereas the 'unsequenced' taxa included mostly Pezizales (including *Helvella* spp.) and Cantharellales (including *Cantharellus* spp.) Within the last 10 years, the DNA of these samples has been extracted from 0.05–10 mg fresh or dried material following one of the five protocols outlined in Suppl. material 1.

For metagenomics approach, we selected 29 vouchered EcM root tip specimens from TU-linked collections of L. Tedersoo and M. Bahram (Table 2). These specimens included either i) rarely occurring EcM fungal lineages not represented by fruitbodies or living cultures (cf. Tedersoo and Smith 2013; n=17), or ii) distinct morphotypes that have remained unamplified and unsequenced in spite of multiple attempts and varying primers (n=12). Samples from the latter category primarily originate from Australia (collected in Tasmania in August, 2006; Tedersoo et al. 2008) and Estonia (collected from various hosts and habitats from May to September, 2013; L. Tedersoo, unpublished). The age of EcM samples ranged from 1.2 to 9.5 years (median, 4.7 years). The DNA of EcM root tips was extracted from fresh or CTAB-stored (100 mM Tris-HCI (pH 8.0), 1.4 M NaCl, 20 mM EDTA, 2% cetyltrimethylammonium bromide) material following one of the four protocols given in Suppl. material 1. In addition, we included two composite samples of soil (AV116 and S160; cf. Tedersoo et al. 2014) and a mock community comprised of 24 fruit-body specimens representing different species (cf. Tedersoo et al. 2015) as controls and for evaluating sequence assembly from more complex samples. Negative controls were not included for sequencing, because of DNA concentration below the detection level.

Molecular techniques

The DNA concentration of all samples was measured using Qubit dsDNA HS Assay Kit (Life Technologies, Carlsbad, CA, USA) and Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA) in January, 2015. Since the DNA concentration of most samples was <1 ng μl^{-1}, the DNA (300 µl) was concentrated up to three times using 750 µl

Table 2. Ectomycorrhiza specimens used for metagenomic sequencing analysis.

Sample code	Identification and EcM lineage	Category	Collection date	Biosample
IO577	Tulasnellaceae, /tulasnella1	Rare	2010-06-xx	SAMN04578193
KP016	Serendipitaceae, /serendipita1	Rare	2011-07-xx	SAMN04578194
L3043d	*Sebacina*[1]	Unseq.	2006-08-xx	SAMN04578195
L3078g	Tulasnellaceae, /tulasnella2	Rare	2006-08-xx	SAMN04578196
L3136g	unidentified	Unseq.[2]	2006-08-xx	SAMN04578197
L3161g	*Discinella*[1]	Unseq.	2006-08-xx	SAMN04578198
L3185g	*Inocybe*[1]	Unseq.	2006-08-xx	SAMN04578199
L3196a	*Discinella*[1]	Unseq.	2006-08-xx	SAMN04578200
L3196g	*Discinella*[1]	Unseq.	2006-08-xx	SAMN04578201
L3273b	Helotiales, /helotiales5	Rare	2006-08-xx	SAMN04578202
L3289	Helotiales, /helotiales4	Rare	2006-08-xx	SAMN04578203
L3371b	Helotiales, /helotiales3	Rare	2006-08-xx	SAMN04578204
L3581g	Helotiales, /helotiales6	Rare	2006-12-xx	SAMN04578205
L3619g	Endogonales, /densospora	Rare	2006-12-xx	SAMN04578206
L7664	Sordariales, /sordariales1	Rare	2010-03-xx	SAMN04578207
L8253	Pyronemataceae, /pyronemataceae1	Rare	2010-07-xx	SAMN04578208
L8574J	*Tomentella*[1]	Unseq.	2013-05-16	SAMN04578209
L8601L	Pyronemataceae, /pyronemataceae2	Rare	2013-06-10	SAMN04578210
L8623J	*Helvella*[1]	Unseq.	2013-06-11	SAMN04578211
L874	Helotiales, /helotiales2	Rare	2005-07-xx	SAMN04578212
L8748B	Helotiales, /helotiales7	Rare	2013-07-03	SAMN04578213
L8760B	Sordariales, /sordariales2	Rare	2013-07-04	SAMN04578214
L8970d	*Tricholoma fulvum*[1]	Unseq.	2013-08-12	SAMN04578215
L9188J	*Tulasnella*[1]	Unseq.	2013-09-20	SAMN04578216
L9238J	*Fischerula macrospora*[1]	Unseq.	2013-09-22	SAMN04578217
L9302J	*Geopora*[1]	Unseq.	2013-10-08	SAMN04578218
N120	Ceratobasidiaceae, /ceratobasidium2	Rare	2008-09-xx	SAMN04578219
TRON3.1	Agaricomycetes, /agaricomycetes1	Rare	2012-04-xx	SAMN04578225
TS1000	Pyronemataceae, /genea-humaria	Rare	2006-08-xx	SAMN04578227

[1]Identification based on ITS sequence from the metagenome.

[2]Unseq., unsequenced.

96% ethanol, 2 µl Pellet Paint Co-Precipitant (cat no 69049–3; Novagen, Madison, WI, USA) and sodium acetate (0.3 M, pH 5.2). DNA precipitation was performed overight at -20 °C. The pellets were washed once with 75% ethanol (-20 °C) and dissolved into MilliQ water, followed by re-determination of the concentration. The obtained 'maximum concentration' ranged from 0.05 to 8.13 ng μl^{-1} (median, 0.57 ng μl^{-1}). All samples were diluted to the concentration of 0.2 ng μl^{-1} (if below, the maximum concentration was used) and 1 ng of DNA was used as an input to prepare sequencing libraries with Nextera XT kit (Illumina Inc., San Diego, CA, USA) according to the instructions of the manufacturer. The concentration of the libraries was measured with Qubit fluorometer and the libraries were pooled equimolarly. The library pools were concentrated with vacuum evaporation and then the library pools were validated by

TapeStation analysis (Agilent Technologies, Santa Clara, USA) and qPCR with Kapa Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA) in order to optimize cluster generation. From each library, 22 pg or 54 pg (dilute samples) of DNA was used in the cluster generation and sequenced on the HiSeq2500 rapid flowcell (Illumina Inc.) with 150 bp paired-end reads protocol.

Bioinformatics

The metagenomics reads of individual samples were demultiplexed and quality-filtered using sdm script of the Lotus pipeline (Hildebrand et al. 2014) with the following options: minAvgQuality=27; maxAmbiguousNT=0; maxHomonucleotide=15; QualWindowWidth=30; QualWindowThreshold=0; TrimWindowWidth=15; TrimWindowThreshold=20. The quality-passed reads were assembled in SPAdes (Bankevich et al. 2012) using default options and kmer sizes 27, 33, 55 and 71. We sought to target the phylogenetically informative genes that are in multiple (nuclear rDNA) or single copies (mitochondrial rDNA, RPB1, RPB2 and TEF1) in the genome. The program sortMeRNA (Kopylova et al. 2012) was used to extract rDNA from raw reads, which were assembled into scaffolds in SPAdes. The fragments were subsequently subjected to bulk blastN search against the entire International Nucleotide Sequence Databases consortium (INSDc) to manually inspect the closest matches focusing on scaffolds of 500–12,000 bases. The coverage of the genomes was esitimated using the Core Eukaryotic Mapping Genes Approach (CEGMA), which gives a genome completeness percentage based on partial and full-length alignments of the target genome with 242 core eukaryotic genes (Parra et al. 2007).

The reference database for genomic and metagenomic fragments comprised 46 fungal genomes and 30 bacterial genomes (present in samples according to rDNA analysis). For the selected SCGs, we used a reference data set of James et al. (2006). Scaffolds containing SCGs were double-checked with manual blastN searches against INSDc and downloaded for trimming and quality evaluation. The sequences of confirmed rDNA genes and SCGs were subjected to multiple sequence alignment using MAFFT 7 (Katoh and Standley 2013) along with 2–5 full-length sequences of the respective genes from Ascomycota and Basidiomycota downloaded from INSDc. The alignments were inspected in SeaView 4 (Gouy et al. 2010) and the flanking non-coding regions were removed. Due to the multiple introns and poor alignability, ca 50–100 bases of flanking regions were retained. For rDNA, we retained the entire copy usually comprising of partial Intergenic Spacer (IGS) 2, SSU, ITS1, 5.8S, LSU and partial IGS1. Due to the poor alignability and multiple introns, mitochondrial rDNA was not trimmed. In many cases, both rDNA and SCGs comprised several different copies that were all kept and submitted to the UNITE database (Abarenkov et al. 2010; accessions UDB028495-UDB028830) and INSDc. The genomic and metagenomic scaffolds of fruit-bodies and root tips were submitted to the Short Read Archive (SRA) of INSDc (Bioproject PRJNA308809; biosample accessions SAMN04578167-SAMN04578254).

Statistics

To evaluate the relative performance of genomics and metagenomics approaches for recovering genetic information of fungi from root tip and fruit-body material of different quality, we constructed linear regression and ANOVA models. First, we tested the effects of the maximum DNA concentration, age of specimen and age of DNA as well as DNA extraction method on the number of reads, size of all scaffolds (confirmed fungal and total and proportion of known fungal) and the longest scaffolds representing rDNA by use of general linear models and forward selection of variables as implemented in Statistica (Statsoft Inc., Tulsa, OK, USA). We determined Pearson correlations among the recovered length of ribosomal and mitochondrial rDNA and SCGs. Further, we arbitrarily chose a threshold of 1500 bases as a criterion for 'successful' sequencing of a barcode, because this value roughly corresponds to the size of mitochondrial SSU and LSU, nuclear SSU and the fragment of commonly amplified nuclear LSU (primers ITS3 and LR5 or LR0R and LR7) as well as SCGs. Differences in sequencing success among markers, sample material (fruit-body vs EcM) and fruit-body type ('old', 'regular' and 'unsequenced', see above) were tested using a series of Fisher's exact tests.

To shed light on the potential issues with DNA secondary structure on amplification and sequencing success in Sanger sequencing, we calculated the minimum free energy (MFE) of the secondary structure of ITS1 and ITS2 reads using RNAstructure (default options for DNA; Reuter and Mathews 2010). The MFE provides an approximation for the stability of a given structure, with lower MFE values indicating more stable structures.

Results

Recovery of genomes

DNA extraction methods yielded similar DNA content and concentration that usually required further concentrating efforts given the small size of our samples. Compared with other methods, the simple ammonium sulphate lysis (cf. Anslan and Tedersoo 2015) retained large amounts of polysaccharides that co-precipitated with DNA, but did not interfere with ligation and sequencing. The HiSeq run produced 553,982,778 individual reads (average length, 145.4 bases), of which 86.7% passed initial quality filtering. Individual genomes and metagenomes were covered by 1,366 to 21,288,678 (median, 5,780,997; SD, 3,863,989) reads with no differences among sample types or fruit-body categories. However, specimen age had a significant negative effect on the recovery of reads in fruit-bodies ($F_{1,54}$=4.4; R^2=0.076; P=0.040) but not in EcM root tips (P>0.1; Fig. 1). The time of DNA isolation had no further impact. The total length of all genomic and metagenomic scaffolds averaged 5.5×10^7 bases (SD, 5.6×10^7) across all samples. It was positively related to the maximum DNA concentration

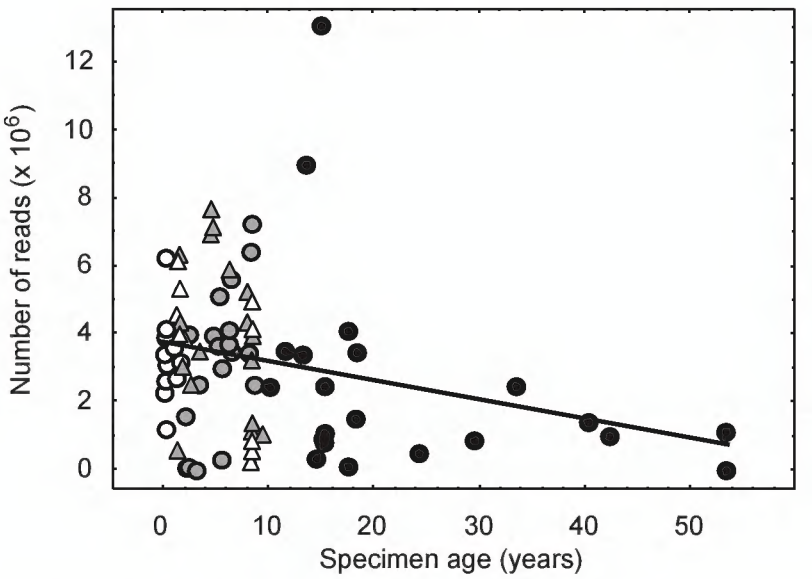

Figure 1. Effect of specimen age on the recovery of reads in the Illumina HiSeq run. Closed circles, 'old' fruit-bodies; shaded circles, 'regular' fruit-bodies; open circles, 'unsequenced' fruit-bodies; shaded triangles, ectomycorrhizal root tips representing unique rare lineages; open triangles, 'unsequenced' ectomycorrhizal root tips.

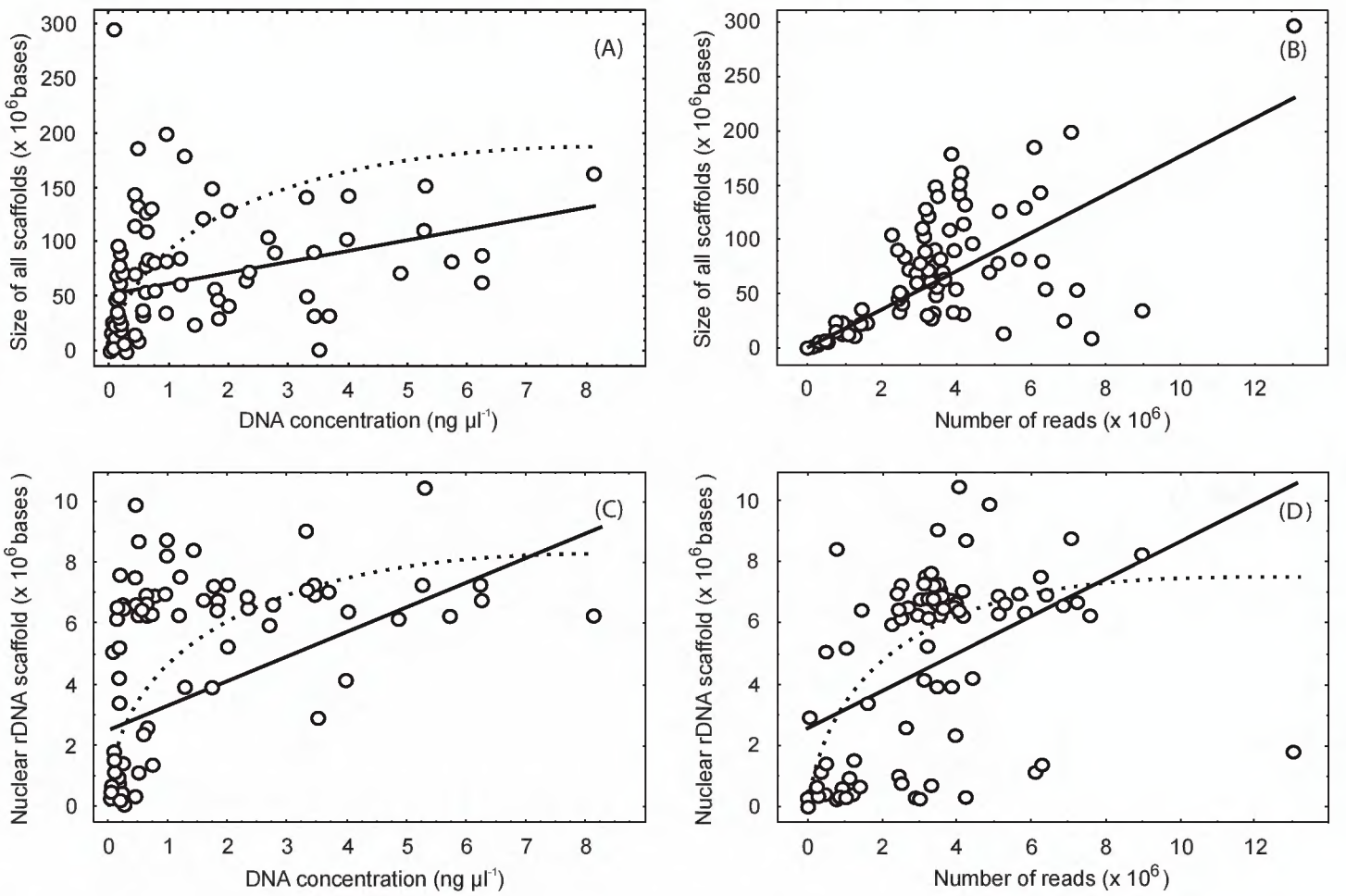

Figure 2. Impact of maximum obtained DNA concentration and number of Illumina HiSeq reads on the size of all scaffolds (**A, B**) and largest nuclear rDNA scaffold (**C, D**). Regular straight lines and dotted lines indicate linear and better fitting logarithmic relationships, respectively.

obtained (partial effect: $F_{1,82}$=10.8; R^2=0.070; P=0.001; Fig. 2A) and total number of reads ($F_{1,82}$=62.1; R^2=0.401; P<0.001; Fig. 2B).

The proportion of genomic and metagenomic sequences belonging strictly to fungi varied greatly across samples, being on average three times lower for EcM root tip

(median, 1.5%; SD, 3.7) compared with fruit-body (median, 4.5%; SD, 15.9) samples ($F_{1,82}$=10.8; R^2=0.098; P=0.001). The lack of closely related reference genomes clearly hampered unequivocal assignment of genomic fragments to fungi or other organisms. Of these, bacteria were the most common organisms in fruit-bodies and EcM root tips, whereas plant scaffolds strongly contributed to the EcM-derived metagenome. However, plant contribution was difficult to establish, because of the large size and ample non-coding regions in plant genomes. Samples of old fruit-bodies and particularly EcM root tips included multiple co-inhabiting fungal species. Their coverage was distinctly lower than that of the target species, but unambiguous separation of these satellite taxa was more difficult for relatively fragmented genomes.

Ribosomal DNA and single copy genes

The coverage of nuclear and mitochondrial DNA and SCGs and their ratio varied greatly across samples independent of sample origin (fruit-body vs. EcM) and category (Suppl. material 1). For the 38 most comprehensively sequenced samples, the standardized ratio of median coverage of nuclear rDNA to mitochondrial rDNA to SCGs was 17.0/4.2/1.0. Notably, *Glomus macrocarpum* (TU116699) and *Endogone* sp. (TU113361) exhibited the corresponding ratios of 1.6/3.6/1.0 and 2.0/2.7/1.0, respectively, indicating low amount of nuclear and mitochondrial rDNA relative to SCGs in their 'fruit-bodies' that are comprised of multinucleate hyphal structures and chlamydospores. In contrast, *Hydnum* sp. (TU116505) and *Cantharellus* sp. (TU116208) stood out as specimens with the highest nuclear rDNA to SCG ratio (43.4/1.0 and 37.6/1.0, respectively).

Among the target regions of fruit-body and EcM samples, nuclear rDNA was relatively more efficiently recovered compared with mitochondrial rDNA and both of these were sequenced with greater success than SCGs (P<0.01 in all cases). There was no difference in the recovery rate among individual SCGs (P>0.5), although RPB1 was completely missing in two samples (ectomycorrhiza of the /genea-humaria lineage TS1000 and *Sarcodon squamosus* TU100663) that exhibited nearly full-length recovery of other SCGs and rDNA. Apart from other taxa, most specimens belonging to Thelephorales contained two highly divergent copies of the TEF1 gene.

The SCGs were significantly less efficiently recovered in EcM samples compared with fruit-body samples (by a factor of 1.9 to 6.2; P<0.001), but the recovery of nuclear and mitochondrial rDNA was comparable between sample types (P>0.1). Across all samples, the maximum DNA concentration (partial effect: $F_{1,82}$=26.7; R^2=0.201; P<0.001; Fig. 2C) and the total number of reads ($F_{1,82}$=23.9; R^2=0.180; P<0.001; Fig. 2D) positively affected the length of the largest nuclear rDNA scaffold. The number of reads necessary to yield full-length rDNA became rapidly saturated at the depth of $2 \times 10^6 - 5 \times 10^6$ sequences for samples with maximum DNA concentration > 0.2 ng μl^{-1} (Fig. 2D).

Fruit-body samples

Within fruit-body collections, rDNA and SCGs were better recovered from 'regular' and recent 'unsequenced' collections than 'old' material (Suppl. material 1). Across all samples, the length of largest scaffolds of nuclear rDNA was strongly correlated to that of mitochondrial rDNA (R=0.724; P<0.001) but not SCGs (P>0.05). The length of largest scaffolds was correlated among all SCGs (0.584<R<0.649; P<0.001).

Fruit-body samples displayed great variation in genomic sequencing success. The 'old' samples sequenced most poorly - i.e., nuclear rDNA >1500 bases could be retrieved only for 43.0% of specimens, which is significantly less compared with 'unsequenced' (73.3%) and 'regular' (84.2%) specimens (P<0.01). Mitochondrial rDNA and SCGs were also relatively poorly recovered in 'old' collections, although the differences were less pronounced among the categories (0.01<P<0.15).

The HTS approach highlighted that primer bias and atypically long ITS markers may account for the Sanger sequencing problems in 'unsequenced' fruit-body samples. In particular, several *Helvella* spp. and *Cantharellus* spp. exhibited ITS1 markers of 500-600 bases that exceed the average values three-fold (Tedersoo et al. 2015). In addition, most *Cantharellus* spp. displayed a 3' terminal mismatch or several mismatches to 'universal' and 'fungal' primers (ITSOF, ITS3, ITS4, LR0R). Besides the regular rDNA copy, *Endogone* sp. (TU116680) exhibited two additional copies that were only 87.2% and 77.5% similar in the ITS region and displayed multiple indels and substitutions in the flanking 5.8S and LSU regions including the highly conserved parts. *Lenzitopsis oxycedri* (UK146) possessed one such abnormal copy with 89.8% ITS sequence similarity, whereas *Glomus macrocarpum* (TU116699) had an extra rDNA copy with 96.0% ITS similarity but no mutations in the flanking 5.8S and LSU fragments. These extra copies had 1.9-2.6 times less coverage than the corresponding regular copies, except that of *L. oxycedri* (54.9-fold difference). The potential problems with sequencing *Clavulina* sp. (TU116528) and *Hydnum* sp. (TU116505) could not be tackled, although the former specimen was 'contaminated' by the DNA of Diptera larvae and a chytrid. The secondary structure of recovered ITS1 and ITS2 sequences had a similar minimum free energy (MFE) and MFE per base in the 'unsequenced' and other categories (Suppl. material 1).

While most collections of stipitate fruit-bodies were relatively free from co-colonization by other fungi, specimens of *Helvella* and those with hypogeous and resupinate fruit-bodies were commonly inhabited by multiple putatively saprotrophic or mycoparasitic fungal taxa. Of these, *Tulasnella*, *Rhizoctonia* (syn. *Ceratobasidium*) and unidentified genera of Eurotiales and Sordariales were the most common. Their nuclear rDNA scaffolds were of relatively lower coverage even if the sequences were nearly full-length. Similar patterns but notably shorter satellite sequences were evident in mitochondrial rDNA (up to, 2000 bases) and SCGs (up to 500 bases).

EcM root tip samples

There were no differences in rDNA and SCG recovery among EcM root tip samples that failed determination previously and those representing rare lineages. Out of 12 previously unidentified EcM root tip samples, only one (L3136g) remained further without identification due to low maximum DNA concentration (0.07 ng/μl) and hence low number of retrieved sequences (173,554 reads). Based on the ITS region, the Tasmanian sequences were identified as *Sebacina* sp. (L3043d), *Inocybe australiensis* (L3185g), and *Discinella* sp. (/helotiales4 lineage; L3161g, L3196a, L3196g). The Estonian sequences were identified as *Tomentella* sp. (L8574J), *Helvella* sp. (L8623J), *Geopora* sp. (L9302J), *Tulasnella* sp. (L9188J), *Tricholoma fulvum* (L8970d) and *Fischerula macrospora* (L9238J) based on the full or partial ITS sequences (Suppl. material 1). The DNA of most EcM samples was apparently degraded, because no primer mismatches, excessively long barcodes, paralogues or deviations in the minimum free energy were evident. Only the *Tulasnella* sp. sample (L9188J) exhibited two mismatches to the ITS3 primer and members of the /helotiales4 lineage had ca. 500-base intron between the ITSOF and ITS1 primer sites.

Using the metagenomics approach, three out of 17 EcM root tips with successful Sanger sequences (L848, L8601, L8760b) failed to retrieve high-quality nuclear rDNA sequences >1500 bases. A single EcM fungus always dominated in nuclear rDNA, but the samples were often co-inhabited by a myriad of ascomycetes, in particular Helotiales, Sordariales, Hypocreales and Dothideales. Basidiomycetes were less common, although *Tulasnella*, Ceratobasidiaceae and Tremellales (*Cryptococcus*) occurred in multiple samples. The ratio of plant to fungal nuclear rDNA varied nearly 80-fold, ranging from 0.21 to 16.3 (median, 1.76) with no apparent differences among host taxa.

Across all 29 EcM root tip metagenomes, fungal TEF1, RPB1 and RPB2 scaffolds >1500 bases were successfully obtained for two, five and fourteen samples, respectively. For 13 samples, none of these SCGs were recovered (scaffolds <500 bases). In successfully sequenced EcM samples, individual SCGs typically occurred in several scaffolds located tens to a few hundred bases apart based on mapping to the alignment. BlastN searches against INSDc and comparisons with rDNA revealed that the largest scaffolds obviously belong to the targeted mycobiont. The co-occurrence of other fungi rendered the taxonomic assignment of SCG scaffolds ambiguous.

Soil and mock community samples

The two highly complex soil metagenomes comprised altogether four fungal nuclear rDNA scaffolds >500 bases in size, three of which were obvious chimeras. The mock community sample included 25 scaffolds encompassing ITS or any of the nuclear rDNA genes (>500 bases). Comparisons with respective Sanger sequences revealed that 32% of these sequences were chimeric, some of which comprising >2 parents. Two of the chimeric sequences were 'circular', i.e. comprised of a full-length rDNA

and fragments of another taxon in one of the ends. Most of the chimeric breaks were located in the conserved regions of 3' half of the SSU and 5' end of LSU, but none were evident in the 5.8S rRNA gene. SSU and LSU of certain congeneric taxa (*Lyophyllum* spp., *Tomentella* spp.) were represented by a consensus sequence that matched perfectly to none of the ingredient specimens. In scaffolds with lower coverage, 5' or 3' ends were sometimes highly diverged from the corresponding Sanger sequence or any database sequences, indicating that artefactual sequences are, to some extent, generated by metagenomics methods.

Discussion

Genomic fragments

We recovered partial fungal genomes and metagenomes from <1 ng DNA of fruit-body and EcM root tip samples with variable success, depending on specimen age and DNA quality (see below). This indicates that fungal genomes can be sequenced from minute amounts of DNA if sufficient quality is secured. The current genome sequencing protocols in the 1000 Fungal Genomes project require three to four orders of magnitude more DNA (http://genome.jgi.doe.gov/programs/fungi/1000fungalgenomes.jsf) that cannot be obtained from tiny samples. In comparison, the genomes of prokaryotes are on average ten times smaller and these have been successfully recovered from common species (upwards 1% relative abundance) in the complex environmental material (Wrighton et al. 2012), multiple single cells (Rodrigue et al. 2009), and high-quality starting material of <0.01 ng DNA (Adey et al. 2010). Because of a single DNA molecule and low proportion of repeats and other non-coding regions, bacterial genomes are easier to assemble compared with eukaryotes that tend to possess long non-coding regions, multiple chromosomes and usually one or two organelles. In our study, taxonomic affinity of especially short scaffolds remained undetermined at the kingdom level based on *de novo* assembly. The paucity of closely related fungal reference material, multiple co-inhabiting organisms and moderate sequencing depth complicated scaffold assembly and rendered estimates of genome size and coverage unreliable (not shown).

Our study aimed to recover the most important genetic markers used for barcoding and phylogenetic reconstruction. Nuclear and mitochondrial rDNA sequences were successfully recovered from most fresh and high-quality samples but typically not from fruit-body specimens >10 years old. For these old specimens, the maximum obtained DNA concentration, a proxy for DNA quality and quantity, remained <0.2 ng/μl. Although other DNA samples were further diluted to this level for library preparation, barcoding markers could not be usually obtained from samples with 0.05-0.2 ng/μl maximum DNA concentration. Because Nextera approach uses DNA fragmentation and 12 cycles of PCR in the ligation step ('tagmentation'), the short DNA molecules of degraded material (Allentoft et al. 2012) may have become over-fragmented or poorly amplified and thus lost from further analytical procedures. This speculation is sup-

ported by 2.3-fold lower yield of reads and 2.8-fold lower proportion of known fungi in 'old' samples compared with 'regular' and 'unsequenced' samples taken together. An 18-year old specimen of *Tomentella ferruginea* (TAAM 166877) represented the oldest collection that was successfully sequenced for the full-length of all rDNA genes and SCGs. In comparison, Staats et al. (2013) successfully sequenced the genome of *Pleurotus ostreatus* fruit-body specimen collected in 1931 by taking advantage of 8000-fold greater amount of DNA and relatively clean vegetative material from the interior of a sporocarp. Old fruit-body samples with large initial amounts of degraded DNA can be prepared for Illumina sequencing using fragmentation-free ligation methods (Carpenter et al. 2013).

Across all samples, nuclear and mitochondrial rDNA were more efficiently recovered compared with SCGs, which reflects the results from amplicon sequencing (Schoch et al. 2012) and scaffold coverage. The range of sequence coverage ratio of nuclear rDNA to SCGs (1.6 to 43.4) is somewhat lower than the previously reported rDNA copy numbers based on qPCR (range, 20 to 200; reviewed in Baldrian et al. 2013). Our indirect estimates should be viewed with caution, because the coverage ratio is based on only 2-3 SCGs and does not account for the AT/GC bias (Perisin et al. 2016). The relative amount of mitochondrial DNA certainly depends on the metabolic activity of a fungus, potentially varying between living cultures, fruit-bodies, EcM root tips and natural mycelium. Taken together, our analyses indicate that fungal species exhibit marked differences in the relative amount of nuclear and mitochondrial rDNA that may further affect metabarcoding- and metagenomics-based estimates of diversity. These results explain the relatively low abundance of Glomeromycota in the soil nuclear rDNA pool (Saks et al. 2014; Tedersoo et al. 2014) and support utilization of SCGs as additional barcodes (e.g. Stockinger et al. 2014).

We sought to uncover the causes why certain fungal species and EcM morphotypes have remained unidentified using direct Sanger sequencing of amplicons. We showed that EcM root tip DNA was degraded and/or comprised of multiple fungal species, which may have disabled direct Sanger sequencing. In fruit-body samples, excessive length of ITS1 sequence might have caused low amplification success in several *Cantharellus* spp. and *Helvella* spp. Due to rapid evolution of rDNA genes in *Cantharellus* (Moncalvo et al. 2006), several otherwise conserved primer sites had one or more mismatches to the templates in the commonly used fungal or eukaryote primers. Furthermore, *Lenzitopsis oxycedri*, *Endogone* sp. and *Glomus macrocarpum* possessed several divergent copies of rDNA that is previously known for a small group of Gloméraceae (Stockinger et al. 2010) and is attributed to the multinucleate habit in that group. Potential ITS paralogues with multiple mutations in the conserved region were evident for the two former species, confirming previous implications based on Sanger sequencing of cloned amplicons (Simon and Weiss 2008) and 454 pyrosequencing of amplicons (Tedersoo et al. 2010; Lindner et al. 2013).

Ribosomal DNA scaffolds from soil and mock community metagenomes indicated artificial generation of a high proportion of chimeric scaffolds during DNA assembly. This demonstrates that markers with long conserved regions such as nuclear

rDNA cannot be reliably assembled even in simple fungal communities. Furthermore, artificial consensus sequences were generated for closely related species with nearly identical SSU and LSU. While such artefacts can be relatively easily tracked in mock communities, metagenomic assembly of rDNA is particularly problematic for natural samples from more complex substrates that comprise hundreds to thousands of fungal species. Due to short scaffolds and the paucity of reference data, we cannot estimate the reliability of scaffold assembly in mitochondrial genes and SCGs, but this may be more problematic with closely related species. Such assembly problems are considered of minor importance in prokaryote metagenomes (Wrighton et al. 2012; Parks et al. 2015) because of a single circular chromosome, lower proportion of repeats, more rapid evolution and more relaxed definition of species/OTUs at 97% SSU similarity (Mende et al. 2013).

Conclusions and perspectives

Taxonomically informative rDNA genes and SCGs can be sequenced from <1 ng DNA of fruit-body and EcM root tip specimens using genomics and metagenomics approaches, respectively. However, fruit-body specimens >10 years old need specific care for obtaining high-quality DNA or require fragmentation-free options for ligation. HTS methods also enabled us to recover large fragments of fungal genomes for a majority of EcM root tips and fruit-bodies that could not be sequenced using Sanger method or that represented unique (including type) material. For high-quality DNA samples, two million (meta)genomic reads were sufficient to recover the full-length nuclear rDNA. Recovery of SCGs was more unpredictable among samples, requiring roughly 10 million unpaired reads. This enables sequencing of ca. 50 fungal genomes on a single 2x150 paired-end Illumina HiSeq run at low coverage (5-10 ×; cf. Stajich 2014). As of January, 2016, a commercial Illumina HiSeq run (5.5×10^8 reads) cost between 5000 and 7000 EUR. However, all individual samples need to be separately ligated with a cost 50-100 EUR sample^{-1}. Thus, the cost per low-coverage fungal draft genome amounts ca. 150-250 EUR. We believe that such low-coverage genomics analyses represent a feasible option to generate multi-gene phylogenomic data sets for tens to hundreds of specimens or mining for the presence and diversity of certain gene families such as carbohydrate active enzymes (CAZymes), antibiotics resistance genes and unique metabolic pathways, but not for routine identification. Targeted enrichment using biotin-linked DNA/RNA probes enables even greater throughput and direct focus on selected markers (Carpenter et al. 2013; Moriarty Lemmon and Lemmon 2013; Manoharan et al. 2015). The full metagenome data also enable to construct draft genomes of prokaryotes and viruses associated with the fruit-body 'mycosphere' and soil 'mycorrhizosphere' that shed light on putative functions and metabolic pathways of these co-occurring microorganisms.

The currently available sequence length and error rate combination does not allow reliable large-scale assembly of genetic information of eukaryotes from complex communities using a single HTS platform. Besides tens and hundreds of millions of Illumina HiSeq

reads, metagenomics analyses would benefit from additional low-coverage sequence analysis of long (up to 3000 bases at 5-8 times circular coverage) fragments as routinely implemented by Pacific Biosciences for in-depth genomic reconstructions. Long amplicon-free backbone sequences reduce the incidence of chimeras and assembly artefacts. Combined with targeted marker capture, this approach would allow greater throughput of eukaryote target genes and more efficient utilization of phylogenetics tools in metabarcoding and community-level functional metagenomic analyses.

Author contribution

LT planned and designed research; UK provided material; IL and PAK performed laboratory analyses; MB, LT and SA analysed data; LT wrote the manuscript with others' input.

Acknowledgements

This work is funded from the Estonian Science Foundation grants 9286, 171PUT, and EMP265. We thank I. Saar and K. Pärtel for providing some of the specimens, DNA extracts and associated metadata. We are grateful to four referees for their constructive comments on earlier versions of the manuscript.

References

Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing B, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U (2010) The UNITE database for molecular identification of fungi – recent updates and future perspectives. New Phytologist 186: 281–285. doi: 10.1111/j.1469-8137.2009.03160.x

Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse M, MacKenzie AP, Caruccio NP, Zhang X, Shendure J (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biology 11: R119. doi: 10.1186/gb-2010-11-12-r119

Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, Campos PF, Samaniego JA, Gilbert TP, Willerslev E, Zhang G, Scofield RP, Holdaway RN, Bunce M (2012) The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proceedings of the Royal Society B 279: 4724–4733. doi: 10.1098/rspb.2012.1745

Anslan S, Tedersoo L (2015) Performance of cytochrome c oxidase subunit I (COI), ribosomal DNA Large Subunit (LSU) and Internal Transcribed Spacer 2 (ITS2) in DNA barcoding of Collembola. European Journal of Soil Biology 69: 1–7. doi: 10.1016/j.ejsobi.2015.04.001

Baldrian P, Vetrovsky T, Cajthaml T, Dobiasova P, Petrankova M, Snajdr J, Eichlerova I (2013) Estimation of fungal biomass in forest litter and soil. Fungal Ecology 6: 1–11. doi: 10.1016/j.funeco.2012.10.002

Bankevich A, Nurk S, Antipov S, Gurevich AA, Dvorkin M, Kulikov AS (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19: 455–477. doi: 10.1089/cmb.2012.0021

Blackwell M (2011) The Fungi: 1, 2, 3 … 5.1 million species? American Journal of Botany 98: 426–428. doi: 10.3732/ajb.1000298

Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft M, Sikora M (2015) The American Journal of Human Genetics 93: 852–864. doi: 10.1016/j.ajhg.2013.10.002

Capella-Gutierrez S, Marcet-Houben M, Gabaldon T (2012) Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. BMC Biology 10: 47. doi: 10.1186/1741-7007-10-47

Dentinger BTM, Gaya E, O'Brien H, Suz LM, Lachlan R, Diaz-Valderrama JR, Koch RA, Aime MC (2016) Tales from the crypt: genome mining from fungarium specimens improves resolution of the mushroom tree of life. Biological Journal of the Linnean Society 117: 11–32. doi: 10.1111/bij.12553

Federhen S (2014) Type material in the NCBI Taxonomy Database. Nucleic Acids Research 2014: 1–13.

Gardes M, White TJ, Fortin JA, Bruns TD, Taylor JW (1991) Identification of indigenous and introduced symbiotic fungi in ectomycorrhizae by amplification of nuclear and ribosomal DNA. Canadian Journal of Botany 69: 180–190. doi: 10.1139/b91-026

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27: 221–224. doi: 10.1093/molbev/msp259

Gueho EP, Kurtzman CP, Peterson SW (1989) Evolutionary affinities of heterobasidiomycetous yeasts estimated from 18S and 25S ribosomal RNA sequence divergence. Systematic and Applied Microbiology 12: 230–236. doi: 10.1016/S0723-2020(89)80067-0

Guidot A, Lumini E, Debaud J-C, Marmeisse R (1999) The nuclear ribosomal DNA intergenic spacer as a target sequence to study intraspecific diversity of the ectomycorrhizal basidiomycete Hebeloma cylindrosporum directly on Pinus root systems. Applied and Environmental Microbiology 65: 903–909.

Hildebrand F, Tadeo R, Voigt AY, Bork P, Raes J (2014) LotuS: an efficient and user-friendly OTU processing pipeline. Microbiome 2: 30. doi: 10.1186/2049-2618-2-30

James TY, Kauff F, Schoch CL (2006) Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443: 818–822. doi: 10.1038/nature05110

Jones MDM, Forn I, Gadelha C, Egan MJ, Bass D, Massana R, Richards TA (2011) Discovery of novel intermediate forms redefines the fungal tree of life. Nature 474: 200–203. doi: 10.1038/nature09984

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30: 772–780. doi: 10.1093/molbev/mst010

Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, Canbäck B, Tunlid A, Grigoriev IV, Hibbett DS, Martin F (2015) Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualisms. Nature Genetics 47: 410–415. doi: 10.1038/ng.3223

Kopylova E, Noé L, Touzet H (2012) SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28: 3211–3217. doi: 10.1093/bioinformatics/bts611

Kõljalg U, Larsson K-H, Abarenkov K, Nilsson RH, Alexander IJ, Eberhardt U (2005) UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytologist 166: 1063–1068. doi: 10.1111/j.1469-8137.2005.01376.x

Lindner DL, Carlsen T, Nilsson RH, Davey M, Schumacher T, Kauserud H (2013) Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. Ecology and Evolution 3: 1751–1764. doi: 10.1002/ece3.586

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA (2010) Population genomics of domestic and wild yeasts. Nature 458: 337–341. doi: 10.1038/nature07743

Manoharan L, Kushwaha SK, Hedlund K, Ahren D (2015) Captured metagenomics: large-scale targeting of genes based on 'sequence capture' reveals functional diversity in soils. DNA Research 22(6): 451–460. doi: 10.1093/dnares/dsv026

Martin F, Aerts A, Ahren A, Brun A, Danchin EGJ, Dochaussoy F (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature 452: 88–92. doi: 10.1038/nature06556

Mende DR, Sunagawa S, Zeller G, Bork P (2013) Accurate and universal delineation of prokaryotic species. Nature Methods 10: 881–884. doi: 10.1038/nmeth.2575

Moncalvo J-M, Nilsson RH, Koster B (2006) The cantherelloid clade: dealing with incongruent gene trees and phylogenetic reconstruction methods. Mycologia 98: 937–948. doi: 10.3852/mycologia.98.6.937

Moriarty Lemmon E, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. Annual Reviews in Ecology, Evolution and Systematics 44: 99–121. doi: 10.1146/annurev-ecolsys-110512-135822

Nguyen NH, Landeroz F, Garibay-Oriel R, Hansen K, Vellinga EC (2013) The *Helvella lacunosa* species complex in western North America: cryptic species, misapplied names and parasites. Mycologia 105: 1275–1286. doi: 10.3852/12-391

O'Brien HE, Parrent JL, Jackson JA, Moncalvo J-M, Vilgalys R (2005) Fungal community analysis by large-scale sequencing of environmental samples. Applied and Environmental Microbiology 71: 5544–5550. doi: 10.1128/AEM.71.9.5544-5550.2005

Parks DH, Imelfort M, Skannerton CT, Hugenholtz P, Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research 25: 1043–1055. doi: 10.1101/gr.186072.114

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061–1067. doi: 10.1093/bioinformatics/btm071

Perisin M, Vetter M, Gilbert JA, Bergelson J (2016) 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. The ISME Journal 10: 1020–102. doi: 10.1038/ismej.2015.161

Pääbo S, Poinar H, Serre D, Jaenicke-Despres, Hebler J, Rohland N (2004) Genetic analyses from ancient DNA. Annual Reviews in Genetics 38: 645–679. doi: 10.1146/annurev.genet.37.110801.143214

Reuter JS, Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11: 129. doi: 10.1186/1471-2105-11-129

Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR (2009) Whole Genome Amplification and De novo Assembly of Single Bacterial Cells. PLoS ONE 4: e6864. doi: 10.1371/journal.pone.0006864

Saks Ü, Davison J, Öpik M, Vasar M, Moora M, Zobel M (2014) Root-colonizing and soil-borne communities of arbuscular mycorrhizal fungi in a temperate forest understorey. Botany 92: 277–285. doi: 10.1139/cjb-2013-0058

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences USA 109: 6241–6246. doi: 10.1073/pnas.1117018109

Simon UK, Weiss M (2008) Intragenomic variation of fungal ribosomal genes is higher than previously thought. Molecular Biology and Evolution 25: 2251–2254. doi: 10.1093/molbev/msn188

Staats M, Erkens RHJ, van de Vossenberg B, Wieringa JJ, Kraijeveld K, Stielow B, Geml J, Richardson JE, Bakker FT (2013) Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. PLoS ONE 8: e69189. doi: 10.1371/journal.pone.0069189

Stajich J (2014) Phylogenomics enabling genome-based mycology. The Mycota 7B: 279–294.

Stockinger H, Krüger M, Schüssler A (2010) DNA barcoding of arbuscular mycorrhizal fungi. New Phytologist 187: 461–474. doi: 10.1111/j.1469-8137.2010.03262.x

Stockinger H, Peyret-Guzzon M, Koegel S, Bouffaud M-L, Redecker D (2014) The largest subunit of RNA polymerase II as a new marker gene to study assemblages of arbuscular mycorrhizal fungi in the field. PLoS ONE 9: e107783. doi: 10.1371/journal.pone.0107783

Tedersoo L, Anslan S, Bahram M, Põlme S, Riit T, Liiv I, Kõljalg U, Kisand V, Nilsson RH, Bork P, Hildebrand F, Abarenkov K (2015) Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MycoKeys 10: 1–43. doi: 10.3897/mycokeys.10.4852

Tedersoo L, Bahram M, Põlme S (2014) Global diversity and geography of soil fungi. Science 234: 1078. doi: 10.1126/science.1256688

Tedersoo L, Jairus T, Horton BM, Abarenkov K, Suvi T, Saar I, Kõljalg U (2008) Strong host preference of ectomycorrhizal fungi in a Tasmanian wet sclerophyll forest as revealed by DNA barcoding and taxon-specific primers. New Phytologist 180: 479–490. doi: 10.1111/j.1469-8137.2008.02561.x

Tedersoo L, Pärtel K, Jairus T, Gates G, Põldmaa K, Tamm H (2009) Ascomycetes associated with ectomycorrhizas: molecular diversity and ecology with particular reference to the Helotiales. Environmental Microbiology 11: 3166–3178. doi: 10.1111/j.1462-2920.2009.02020.x

Tedersoo L, Smith ME (2013) Lineages of ectomycorrhizal fungi revisited: foraging strategies and novel lineages revealed by sequences from belowground. Fungal Biology Reviews 27: 83–99. doi: 10.1016/j.fbr.2013.09.001

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerknoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF (2012) Fermentation, hydrogen, and sulfur Metabolism in multiple uncultivated bacterial phyla. Science 337: 1661–1665. doi: 10.1126/science.1224041

Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, Lanz C, Martin FN, Kamoun S, Krause J, Thines M, Weigel D, Burbano HA (2013) The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. eLife 2: e00731.

Supplementary material 1

Full information and metadata about the genomic and metagenomic samples

Authors: Leho Tedersoo, Ingrid Liiv, Paula Ann Kivistik, Sten Anslan, Urmas Kõljalg, Mohammad Bahram

Data type: table

Explanation note: Detailed information about metadata, DNA quality and genomic/metagenomic results of fruit-body and EcM root tip samples.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.